Effectiveness of Data Augmentation and Pretraining for Improving Neural Headline Generation in Low-Resource Settings

Martinc, Matej

European Language Resources Association (ELRA)
2022-07

Martinc , M , Montariol , S , Pivovarova , L & Zosa , E 2022 , Effectiveness of Data Augmentation and Pretraining for Improving Neural Headline Generation in Low-Resource Settings . in Proceedings of the 13th Language Resources and Evaluation Conference . þÿEuropean Language Resources Association (ELRA) , pp. 3561 3570 , L
Marseille , France , 20/06/2022 . <
http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.381.pdf >

http://hdl.handle.net/10138/346642

Effectiveness of Data Augmentation and Pretraining for Improving Neural Headline Generation in Low-Resource Settings

Matej Martinc[1], Syrielle Montariol[2], Lidia Pivovarova[3], Elaine Zosa[3]
[1]Jozef Stefan Institute, [2]INRIA Paris, [3]University of Helsinki
matej.martinc@ijs.si, syrielle.montariol@inria.fr, first.last@helsinki.fi

Abstract

We tackle the problem of neural headline generation in a low-resource setting, where only limited amount of data is available to train a model. We compare the ideal high-resource scenario on English with results obtained on a smaller subset of the same data and also run experiments on two small news corpora covering low-resource languages, Croatian and Estonian. Two options for headline generation in a multilingual low-resource scenario are investigated: a pretrained multilingual encoder-decoder model and a combination of two pretrained language models, one used as an encoder and the other as a decoder, connected with a cross-attention layer that needs to be trained from scratch. The results show that the first approach outperforms the second one by a large margin. We explore several data augmentation and pretraining strategies in order to improve the performance of both models and show that while we can drastically improve the second approach using these strategies, they have little to no effect on the performance of the pretrained encoder-decoder model. Finally, we propose two new measures for evaluating the performance of the models besides the classic ROUGE scores.

Keywords: Natural language generation, Headline generation, Low resource languages

1. Introduction

Neural approaches for natural language generation (NLG) have mushroomed during past few years. The most common idea is to employ approaches that have shown good performance in machine translation (or another sequence-to-sequence task) and treat the generation task as a translation task between an input text and the generated output text (Wen et al., 2015; Cho et al., 2014). The most popular text generation is automatic summarization, and recent years have seen huge advances in automatic generation of high-quality summaries. The newest approaches, such as BART (Lewis et al., 2020), employ an encoder-decoder transformer architecture (Vaswani et al., 2017), which "translates" the input text into an output summary.

Due to large textual resources required by these NLG systems, research on this topic mostly focused on high-resource languages such as English, since the lack of data makes the training of these approaches from scratch infeasible in some low-resource domains and languages (Gkatzia, 2016). While recently some multilingual models which also cover low-resourced languages (Liu et al., 2020) have been proposed, most low-resource languages still lack efficient monolingual language generation systems. Therefore, to generate texts for these languages with a neural architecture but without large datasets and substantial computational resources—required for extensive pretraining of encoder-decoder models—we are left with two options:

Using a multilingual NLG system that supports the low-resource language in which we wish to generate text. The options are limited here, with the multilingual generation models ProphetNet-Multi (Qi et al., 2021b) and mBART-50 (Tang et al., 2020) currently being the models supporting the most languages (52 and 50 respectively, including some low-resource ones). The possible downside of using this approach is the so-called curse of multilinguality (Conneau et al., 2020), i.e., a trade-off between the number of languages the model supports and the overall decrease in performance on monolingual and cross-lingual benchmarks.

Training a multilingual encoder-decoder NLG system from scratch, with the downside being that the performance of the model will be most likely directly correlated to the amount of available training data. One possible solution to partially circumvent this problem is to employ an approach proposed by Rothe et al. (2020), which relies on the usage of two pretrained transformers, combined into an encoder-decoder NLG architecture. In this case, only the cross-attention layer needs to be trained from scratch, and since the combined model can leverage the knowledge gained during the language model pretraining, it requires less training data for optimal performance, at least in theory. An upside of this approach is that these multilingual pretrained transformer-based language models (Vaswani et al., 2017) have been recently trained for a plethora of low-resource languages[1], meaning that this approach can be used for much more languages than by using a pretrained multilingual NLG system.

While automatic summary generation is very popular, generation tasks, which focus on production of more creative content such as headlines or slogans, receive less focus. However, a headline can also be considered as a sort of summary, since it is a vehicle that carries the most important information about the news article content. The newest approaches for headline generation

[1] The Huggingface library currently offers pretrained transformers for 168 languages: `https://huggingface.co/models`

based on this idea have obtained promising results, but this research is once again mostly focused on English (Shen et al., 2016). On the down side, these approaches are difficult to employ in real-life scenarios due to a special type of overfitting called "hallucination", where the system produces non-factual outputs that are not based on the data presented in the input (Reiter, 2018; Dušek et al., 2019). This severely limits the application of these systems in the domain of newspaper articles, where the production of factual text is essential. These systems also lack interpretability and their evaluation could be unreliable unless conducted manually by humans. It has been shown that commonly used automated evaluation metrics do not necessarily correlate well with human judgement (Reiter and Belz, 2009; Dušek et al., 2018).

We tackle some of the problems and research gaps introduced above. These are our main contributions:

- We address the generation of creative texts, news headlines, in a low-resource multilingual setting with neural encoder-decoder architectures. More specifically, we compare the two distinct approaches for NLG described above. In the first approach, we use a pretrained monolingual NLG system BART (Lewis et al., 2020) or multilingual mBART (Liu et al., 2020) (depending on the language). In the second approach, we train the NLG model from scratch, relying on pretrained BERT models combined into an NLG encoder-decoder, same as in Rothe et al. (2020)[2].

- We explore two techniques for reducing the needed amount of training data, namely data augmentation and domain-specific pretraining. We focus on evaluating how these strategies affect both types of models and conclude that they have a significant influence only in the second approach, where pretrained BERT models are combined into an NLG encoder-decoder.

- We propose two evaluation measures that have not been applied for headlines generation in the literature. Both measures focus on the semantic similarity between correct and generated headlines and therefore complement the established ROUGE score, which measures a word overlap and was criticized in the past for not considering semantic similarity.

- We offer a manual error analysis in order to determine how the proposed data augmentation and pretraining tactics affect both models and to pinpoint mistakes specific for each model.

2. Related Work

As stated above, most recent approaches to headline generation consider it as a summarization task and employ state-of-the-art neural summarization models. These models have been used to tackle several distinct variants of the headline generation task, such as bilingual headline generation (Shen et al., 2018), headlines for community question answering (Higurashi et al., 2018), multiple headline generation (Iwama and Kano, 2019) and also user-specific headline generation used in the recommendation systems (Liu et al., 2018).

Liang et al. (2020) compare multiple text noising strategies for training, showing large improvements on the headline generation task. The best noising strategy consists of sampling a number of token spans from the original text with span lengths drawn from a Poisson distribution, and then replacing each token span with a single [MASK] token.

While most research is still focused on English, recently some multilingual benchmarks for news headline generation were proposed. Among the well-known benchmarks, X-GLUE (Liang et al., 2020) includes a headline generation task, covering 5 high-resource languages (German, English, French, Spanish and Russian) and using BLEU-4 score as the metric. The training dataset contains 300K examples, and development and test datasets contain 10k examples. In this benchmark, XLM-R (Conneau et al., 2020) and M-BERT (Devlin et al., 2019), initialized as encoder-decoder models and fine-tuned on the downstream task, are outperformed by the Unicoder (Huang et al., 2019), a universal language encoder trained to be language-agnostic by being pretrained on cross-lingual tasks.

A more general benchmark for text generation is GLGE[3], including 4 abstractive text summarization tasks, CNN/DailyMail (Hermann et al., 2015) (See et al., 2017), Gigaword (Rush et al., 2015) (Graff et al., 2003), XSum (Narayan et al., 2018), and MSNews. Gigaword and MSNews both use news headlines as targets, while in the other two tasks informative summaries need to be generated. All tasks are in English, and the benchmarks are divided into three versions, from easy to hard. Prophetnet (Qi et al., 2020) and its other version ProphetNet-X (Qi et al., 2021a) beat Unicoder on this second benchmark, but are outperformed by BART (Lewis et al., 2020) on the hard version of the benchmark. ProphetNet and BART were also trained on a multilingual corpus. ProphetNet-Multi is trained on the 101GB Wiki-100 corpus and 1.5TB Common Crawl2 data. Similarly, mBART, which we employ in this study and is described in more detail in Section 3.1, is trained on 25 languages and its bigger version mBART-50 on 50 languages.

[2] The code for experiments is available under the MIT licence at `https://gitlab.com/matej.martinc/low_resource_headline_generation`.

[3] `https://github.com/microsoft/glge`

3. Methodology

3.1. The Models

For our experiments, we use two state-of-the-art summarization systems. The first system is BART (Lewis et al., 2020), a denoising autoencoder for pretraining sequence-to-sequence models[4]. BART employs a standard transformer-based neural machine translation architecture and is pretrained on several denoising tasks, in which the original text is corrupted and the model is trained to generate an uncorrupted output. The training corpus is corrupted by either randomly shuffling the original sentences or by using an in-filling scheme, where spans of text are replaced with a single mask token. BART achieved new state-of-the-art results on a set of tasks, among them classification, question answering, and summarization. We employ BART for English, while for experiments on Estonian and Croatian we use its multilingual version mBART-50 (Tang et al., 2020).

The other approach, proposed in Rothe et al. (2020), relies on a combination of pretrained transformer-based language models. Using one language model as an encoder and the other as a decoder, the authors demonstrate the efficacy of pretrained language models for sequence generation, leading to state-of-the-art results on several tasks, among which machine translation and text summarization. We use as encoders and decoders two pretrained BERT models (Devlin et al., 2019), which are available for all languages covered in our experiments, and name this approach BERT-ED.

The main difference between the two approaches is that BART has already been pretrained as an encoder-decoder model on a large corpus consisting of books and Wikipedia (i.e. the same corpus as BERT), and mBART-50 on a large dataset containing texts from 50 languages extracted from the Common Crawl (CC) (Wenzek et al., 2020). On the other hand, BERT-ED consists of two pretrained BERT models[5] connected by a cross-attention layers, which are *randomly initialized*. We suspect this difference would result in a gap in performance between the two systems when trained on a relatively small corpora in a low-resource setting. We hypothesise that while BART will be harder to adapt for a specific headline generation task due to its extensive pretraining as an encoder-decoder, it would nevertheless return semantically and grammatically better headlines. Since the cross-attention layer in the system composed of two BERT models has not been pretrained, this approach might require more training data to generate semantically and grammatically correct headlines. It would nevertheless be easier to adapt to a specific task and domain at hand.

3.2. Training Schemes

As mentioned above, our main focus is to evaluate these systems in a low-resource setting. Most related work train neural models on large datasets consisting of more than 100,000 documents. In contrast, we test the models in a low-resource setting, on datasets ranging from 10 000 to roughly 30 000 documents, and investigate whether using and combining different pretraining schemes can improve the performance of the model. More specifically, we test three distinct pretraining techniques:

- **Text infilling**: As proposed by Lewis et al. (2020), about 20% of the training corpus is corrupted by an in-filling scheme, where spans of text are replaced with a single mask token. The encoder-decoder is then trained to generate the original text from the corrupted input.
- **Sentence shuffling**: Same as in Lewis et al. (2020), the input sentences are randomly shuffled and the model is trained to generate the original text with the correct sentence order.
- **2 tasks**: The model is first trained to restore the correct order of shuffled sentences and than to restore the corpus corrupted by the text in-filling scheme.

Note that pretraining is performed using only the headline generation training dataset and no additional data is used. This way, we inspect if the model's performance can be improved by extensive pretraining instead of obtaining more data.

3.3. Data Augmentation

To increase the size of the training corpus we employ several data augmentation techniques.

- **BERT-based augmentation**: 20% of the words in the news article are masked. Then, the masked article is fed to the BERT model, who proposes probable candidates for the masked tokens. These tokens are replaced by the most probable candidates, creating new articles to be added to the training set.
- **Word2vec augmentation**: For each news article in the train set, we replace random words in the articles by synonyms proposed by the Word2vec model.[6]

[4] We opted to test this model instead of the alternative ProphetNet-X (Qi et al., 2021a) since it is more comparable in size to the other tested model BERT-ED (see below), making the comparison fairer.

[5] For English we used two "bert-base-uncased" models, for Estonian and Croatian we used the FinEst BERT and CroSloEngual BERT described in (Ulčar and Robnik-Šikonja, 2020), respectively.

[6] We set the number of runs parameter to 5 and probability of replacement to 0.3 (i.e., the algorithm goes through the text five times and tries to augment each sentence with a 0.3 probability). English word2vec embeddings are trained on the Google News dataset, Croatian word2vec embeddings are trained on the Croatian Web Corpus (HrWAC) (Ljubešić and Erjavec, 2011) while the Estonian embeddings are trained on the Estonian Reference Corpus (Kaalep et al., 2010).

- **Wordnet augmentation**: This method is similar to the previous one, but replacement candidates are obtained from Wordnet.

- **EDA augmentation**: EDA, proposed by Wei and Zou (2019), consists of four operations: Wordnet synonym replacement, random insertion, random swap, and random deletion.

- **Mixed augmentation**: Each article in the train set is first augmented with Word2vec. The augmented article is fed to the EDA-based augmentation and the output of this augmentation is additionally fed to the Wordnet-based augmentation.

All augmentation techniques except for BERT have been previously established and are available in the TextAugment library[7]: For English, we used all augmentation strategies. For Croatian and Estonian only BERT and word2vec augmentations are available since Wordnet is not available for these languages.

For each original article in the train set, we generate 5 augmented articles using the algorithms described above. These new articles are inserted into the original training set and used for training of the headline generation model. We opted to generate five augmented texts per article, as initial experiments suggested that using a smaller number results in an insufficient increase of the training dataset, and using a larger number results in repetitions of the training examples.

3.4. Evaluation

For evaluation, we employ the ROUGE score, which is the current standard for evaluating generated summaries and headlines. However, ROUGE score does not necessarily have sufficient correlation with human judges (Reiter and Belz, 2009; Dušek et al., 2018) because it only compares n-gram overlap and therefore does not represent well the semantic similarity between true and generated headlines. To alleviate this problem, we propose two new evaluation measures that consider semantic similarity. The first measure, *semantic similarity* (SS), measures cosine distance (CD) between the embedding of the true and generated headline. We employ sentence transformers (Reimers and Gurevych, 2019) for generating embeddings for true and generated headlines.[8] The second evaluation approach is motivated by Yin et al. (2019), who used a pretrained *natural language inference* (NLI) sequence-pair classifier as a zero-shot text classifier. Considering the true headline as the "premise" and each generated headline as the "hypothesis", we use the NLI model to predict whether the premise entails the hypothesis. We take the probability of the entailment between a true and a generated headline as a measure of headline quality. Note that this measure is only used for English experiments, since there is no available model pretrained for NLI that covers Croatian and Estonian.[9]

4. Experiments

4.1. Experimental Setting

Experiments were conducted on three datasets, namely the Estonian ExM news article dataset (Purver et al., 2021a), the Croatian 24sata news article dataset (Purver et al., 2021b) and the English KPTimes dataset (Gallina et al., 2019). The dataset statistics are presented in Table 1. For Croatian and Estonian, we use the same train and test dataset splits as in the recent study on keyword extraction (Koloski et al., 2021).

The English dataset is included in our experiments to serve as a benchmark for several comparisons. First, we wish to research whether there is a discrepancy in the quality of produced headlines between English (for which most NLG models are originally produced) and two low-resource languages, Estonian and Croatian. Second, besides conducting low-resource experiments, the abundance of resources in English allows us to obtain results for the high-resource scenario, to which we can compare our low-resource results. For this reason, we use both the large KPTimes train set, containing about 260,000 news articles, and the original KPTimes validation set, containing 10,000 articles, which we employ as a 'low-resource' English train set and train models on it. Since we do not use these datasets as training and validation sets, we refer to them as 260K and 10K respectively to avoid terminology confusion.

Both BART and BERT-ED approaches are first tested in a high resource scenario, i.e., by training them on the 260K KPTimes train set. The results of these experiments are used as a reference point of how well these models work in an ideal scenario with plenty of data available, to which we can compare results of our low-resource experiments. Next, both of these models are trained on the 10K set, the Estonian train set, and the Croatian train set without any additional pretraining or data augmentation. These low-resource reference points are used as baselines that we want to improve through various pretraining and data augmentation methods.

In our experiments, we employ the same training and generation regime for both models. The input news articles are truncated at 128 tokens, since we assume that the most important content of the news, to which the title most likely refers to, is covered at the beginning of the article. The length of the output is limited to 30 tokens; finally, for generation we employ a beam search of size 5 and early stopping[10].

[7] `https://github.com/dsfsi/textaugment`

[8] More specifically, we employ the "sentence-transformers/paraphrase-MiniLM-L6-v2" for experiments on English and "sentence-transformers/paraphrase-xlm-r-multilingual-v1" for experiments on Croatian and Estonian. Both models are available in the Huggingface library.

[9] For English, we employ the "typeform/distilbert-base-uncased-mnli" for entailment predictions.

[10] We pretrain all models for 10 epochs per task, using

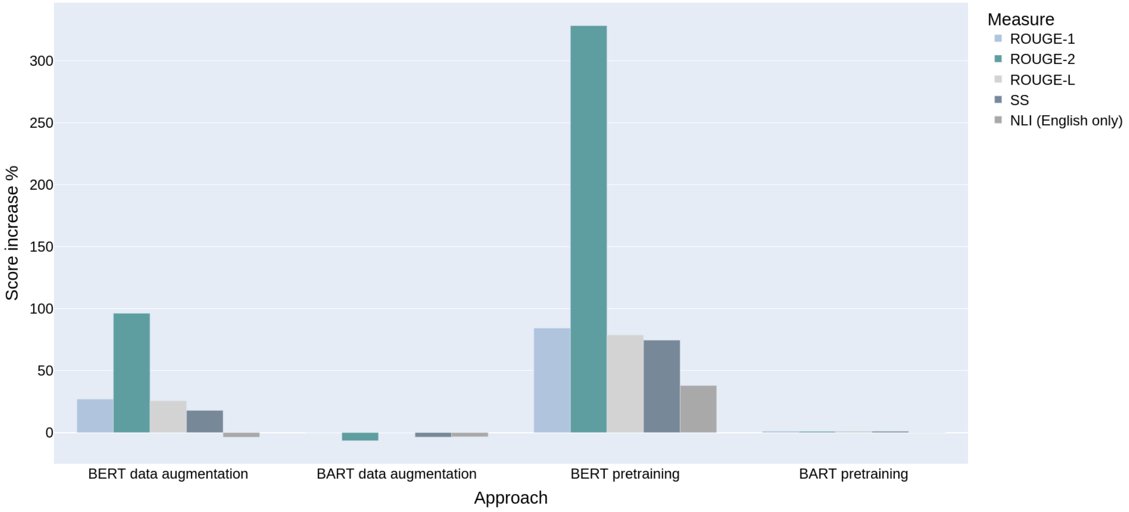

Figure 1: Average *increase* in performance for pretraining and data augmentation approaches for both models across the three languages according to five evaluation measures: three ROUGE scores, semantic similarity (SS) and NLI (only for English).

Language	train set	test set
English 260K (KPTimes train)	259,923	10,000
English 10K (KPTimes valid)	10,000	10,000
Croatian	32,223	3,582
Estonian	10,750	7,747

Table 1: News datasets used for empirical evaluation of headline generation (number of documents).

4.2. Results

The results of the experiments on the English dataset are presented in Table 2 and the results of the experiments on Estonian and Croatian datasets are presented in Table 3[11].

Both BERT-ED and BART models perform well in the ideal high-resource scenario when trained on the large 260K train set (see approaches labeled as “BASELINE 260K” for both English models), with BART outperforming BERT-ED by roughly 4 points according to all three ROUGE scores, by about 2 points according to SS and by almost 5 points according to NLI.

On the other hand, when the models are compared in a low-resource scenario, the gap between the model’s performance drastically increases (see approaches labeled as BASELINE for Estonian and Croatian models and the approach labeled as “BASELINE 10K” for English models). For example, the English BART model trained on the English 10K dataset outperforms BERT-ED trained on the same dataset by about 20 points according to ROUGE-1, by about 10 points according to ROUGE-2 and NLI, by about 16 points according to ROUGE-3, and by about 25 points according to SS. This is due to the drastic decrease in BERT-ED’s performance when trained on the small 10K dataset. Similar phenomena can be observed for the other two languages, Croatian and Estonian, with the performance being especially bad on the Estonian corpus, where the model has trouble converging and achieves very low ROUGE scores.

While the results for BERT-ED clearly indicate that only training the model from scratch on a corpus of limited size is not a viable option, BART-based models on the other hand show more robust performance, even when trained in the low-resource scenario. For English, training the BART model on the 10K dataset results in a modest drop of about 3 points according to all criteria, when compared to the BART model trained on the 260K dataset. The results for Estonian and Croatian are worse, yet still much better than for the BERT-ED-based models. On Estonian, the multilingual mBART model achieves ROUGE-1 of 26.2, ROUGE-2 of 12.3, ROUGE-L of 24.3 and SS score of 56.7.

While comparison of ROUGE and SS scores across languages is problematic,[12] these scores—and the manual inspection confirming the quality of the produced headlines—indicate that an extensively pretrained multilingual model can be successfully applied in a low-

the default learning rate of $2e-5$. The same configuration is employed during the headline generation fine-tuning. The number of epochs was chosen empirically, on the basis of initial experiments that suggested that a more extensive pretraining/fine-tuning could result in overfitting.

[11] Note that we report results for single runs and not for averages across several random seeds due to computational constraints. While results for single runs are indeed less reliable, main claims we make should however not be affected by this shortcoming, since they are based on substantial differences in performance according to all evaluation criteria.

[12] This is especially true when comparison needs to be made between a morphologically rich language, such as Estonian, and a morphologically less diverse language, such as English.

Approach		ROUGE-1		ROUGE-2		ROUGE-L		SS		NLI	
English BERT-ED-based models											
BASELINE	10K	10.2		1.4		9.6		26.7		15.4	
	260K	27.6		10.1		25.1		52.8		32.1	
AUGMENTATION	bert	13.2	3.0	2.3	0.9	12.2	2.6	33.4	6.7	15.7	0.3
	w2v	9.7	-0.5	1.6	0.2	8.9	-0.7	28.3	1.6	14.9	-0.5
	mix	10.4	0.2	1.7	0.3	9.6	0.0	25.6	-1.1	13.1	-2.3
	eda	12.8	2.6	2.2	0.8	11.9	2.3	32.4	5.7	15.2	-0.2
	wordnet	12.4	2.2	2.1	0.7	11.5	1.9	31.8	5.1	15.2	-0.2
PRETRAINING	infilling	11.7	1.5	1.9	0.5	10.7	1.1	33.6	6.9	18.9	3.5
	shuffling	12.9	2.7	2.6	1.2	11.8	2.2	37.5	10.8	18.8	3.4
	2 tasks	16.5	6.3	4.6	3.2	15.1	5.5	43.9	17.2	25.9	10.5
English BART-based models											
BASELINE	10K	**29.0**		**10.9**		**26.0**		52.5		34.1	
	260K	31.9		13.1		28.7		55.0		36.8	
AUGMENTATION	bert	28.5	-0.5	10.5	-0.4	25.6	-0.4	52.4	-0.1	34.0	-0.1
	w2v	27.8	-1.2	10.1	-0.8	25.1	-0.9	51.4	-1.1	32.0	-2.1
	mix	27.7	-1.3	10.2	-0.7	25.0	-1.0	51.0	-1.5	32.2	-1.9
	eda	28.3	-0.7	10.4	-0.5	25.5	-0.5	52.2	-0.3	33.2	-0.9
	wordnet	28.2	-0.8	10.3	-0.6	25.3	-0.7	52.0	-0.5	33.4	-0.7
PRETRAINING	infilling	29.0	0.0	10.9	0.0	26.0	0.0	**52.7**	**0.2**	34.2	0.1
	shuffling	28.8	-0.2	10.8	-0.1	25.9	-0.1	52.5	0.0	**34.3**	**0.2**
	2 tasks	28.7	-0.3	10.7	-0.2	25.9	-0.1	52.4	-0.1	34.1	0.0

Table 2: Results of experiments on the English datasets. Best results in a low resource setting (i.e., excluding the BART and BERT-ED models trained on English 260K dataset) per evaluation measure are **bolded**. For each measure, we report its absolute value (the first number) and the difference with the baseline model (the second, colored, number). Since all experiments with data augmentation and pretraining are run on the 10K dataset, differences are computed respectively to the 10K baseline, i.e. the first row of results for each model.

resource scenario. The mBART results for Croatian are worse, which is interesting, since the Croatian train set is three times the size of the Estonian one. They can nevertheless be explained by the fact that mBART-50 was pretrained on a much smaller Croatian corpus than the Estonian one (Tang et al., 2020).

Next, we discuss the results of the **data augmentation** and pretraining experiments. Generally speaking, the results indicate that these experiments have on the one hand a significant influence on the performance of BERT-ED-based models and a negligible influence on the performance of the BART-based models. When it comes to English data augmentation, all but one (Word2Vec augmentation) method manage to beat the BERT-ED 10K baseline score. The biggest improvement can be observed for the BERT augmentation. Decent improvements according to all criteria can also be observed when EDA and Wordnet augmentation are used. Mix augmentation does not work that well, probably because texts become very different after the multi-step process and not always preserve the original meaning. It is hard to fine-tune augmentation parameters, since this would require retraining of the corresponding headline generation model.

For Croatian, the data augmentation improvements are smaller than for English; BERT data augmentation does not work at all. As the Croatian training dataset is three times bigger than the English and Estonian ones, we deduce that increasing the dataset size with data augmentation techniques might be less beneficial for larger datasets. The highest improvements over the BERT-ED baseline for data augmentation are observed for the Estonian dataset. Indeed, the BERT-ED baseline—which most likely did not converge due to the lack of training data—returns mostly repetitive or empty strings, while data augmentations apparently creates enough additional training data to generate more coherent content.

For the BART-based models, all data augmentation strategies lead to scores lower than the baseline for all languages. While the reduction is in most cases minimal, these scores nevertheless do indicate that the augmented data is not of sufficient quality for the pretrained model to obtain useful information that can be successfully leveraged during NLG training.

By **pretraining** the BERT-ED-based models, using text infilling and sentence shuffling tasks, on the same datasets on which they are later fine-tuned for headline generation, we obtain substantial performance boosts. The increase in performance is even larger than with data augmentation. For English and Estonian, it is especially useful to apply both pretraining regimes, sentence shuffling and text infilling, sequentially (see the row in Tables 2 and 3 labeled as "PRETRAINING 2 tasks"). For Croatian, text infilling works slightly better than sentence shuffling according to most criteria, but combining these two approaches does not improve the performance.

Approach		ROUGE-1		ROUGE-2		ROUGE-L		SS	
Croatian BERT-ED-based models									
BASELINE		9.6		1.0		8.9		29.7	
AUGMENTATION	bert	2.5	-7.1	0.0	-1.0	2.5	-6.4	10.2	-19.5
	w2v	11.0	1.4	1.4	0.4	0.1	1.1	33.7	4.0
PRETRAINING	infilling	16.6	7.0	4.2	3.2	14.8	5.9	44.9	15.2
	shuffling	15.2	5.6	3.6	2.6	13.4	4.5	43.9	14.2
	2 tasks	15.4	5.8	4.2	3.2	13.6	4.7	45.9	16.2
Croatian BART-based models									
BASELINE		20.5		7.3		18.1		49.6	
AUGMENTATION	bert	19.8	-0.7	6.8	-0.5	17.6	-0.5	49.8	0.2
	w2v	18.3	-2.2	5.8	-1.5	16.3	-1.8	47.9	-1.7
PRETRAINING	infilling	21.0	0.5	**7.5**	**0.2**	18.6	0.5	**51.1**	**1.5**
	shuffling	**21.2**	**0.7**	7.4	0.1	**18.7**	**0.6**	50.8	1.2
	2 tasks	20.8	0.3	7.2	-0.1	18.4	0.3	50.9	1.3
Estonian BERT-ED-based models									
BASELINE		3.9		0.3		3.8		17.9	
AUGMENTATION	bert	9.8	5.9	2.5	2.2	9.4	5.6	36.9	19.0
	w2v	8.5	4.6	2.1	1.8	8.1	4.3	34.4	16.5
PRETRAINING	infilling	13.9	0.1	4.3	4.0	13.2	9.4	44.0	26.1
	shuffling	11.3	7.4	2.8	2.5	10.7	6.9	40.7	22.8
	2 tasks	17.6	13.7	6.5	6.2	16.3	12.5	49.8	31.9
Estonian BART-based models									
BASELINE		26.2		12.3		24.4		56.7	
AUGMENTATION	bert	25.4	-0.8	11.6	-0.7	23.8	-0.6	55.9	-0.8
	w2v	23.0	-3.2	9.8	-2.5	21.5	-2.9	53.5	-3.2
PRETRAINING	infilling	**27.1**	**0.9**	**12.9**	**0.6**	**25.2**	**0.8**	**57.2**	**0.5**
	shuffling	26.6	0.4	12.6	0.3	24.8	0.4	56.9	0.2
	2 tasks	26.6	0.4	12.3	0.0	24.6	0.2	56.6	-0.1

Table 3: Results of experiments on the Croatian and Estonian datasets. Best results per language and per evaluation measure are **bolded**. For each measure, we report its absolute value (the first number) and the difference with the baseline model (the second, colored, number). The differences are computed in respect to the baseline.

Pretraining the BART-based models leads to small improvements for Estonian and Croatian, and to small reduction for English. The monolingual English BART, which was extensively pretrained on a massive English corpus using the same denoising tasks we employ here, apparently does not profit from the additional pretraining on a small corpus. The pretraining experiments for the multilingual mBART-50 on the other hand consistently show small improvements across all three pretraining regimes and for both languages.

The average increase in performance for data augmentation and pretraining across all languages and for both models is visualized in Figure 1. It is visible that the employment of data augmentation or pretraining leads to on average much larger increase in performance when BERT-ED-based models are used. The measure that benefits the most from these additional steps is ROUGE-2, most likely since this is the hardest criterion of the model's quality, which is only slightly above zero for most baseline BERT-ED-based approaches. On the other hand, the figure clearly shows that both pretraining and data augmentation have only a marginal effect on the BART-based models.

5. Qualitative results

We manually checked the outputs of several English models. The BART model, fine-tuned on the 10K dataset produces one of the the best results. However, it can hallucinate (see Example 2 in Table 4) or shift the focus of the headline. The manual inspection did not reveal any large differences between the BART-based model trained on the 10K dataset and on the 260K dataset. Interestingly, Example 1 results in identical outputs for BART models trained on both datasets, as well as in *all* other modifications we try with BART. Variation between outputs are rare and, in most cases, not significant; thus, it is hard to judge which outputted headline is better. On the contrary, the performance of the BERT-ED-based model trained on the 10K dataset drops drammatically compared to the one trained on the 260K dataset, as could be seen in the same table. In most cases, it produces ungrammatical sequences with many repetitions.

Data augmentation only slightly improves the perfor-

	Example 1	Example 2
True headline	martial law is rescinded in a philippine province	fighting n. y. c. soda ban, industry focuses on personal choice
BART 260K	philippine president lifts martial law	soda industry fights new york city's soda ban
BART 10K	philippine president lifts martial law	soft-drink industry takes aim at sugary drinks
BERT-ED 260K	philippine president lifts martial law in southern philippines	soft - drink industry seeks to fight sugary drinks ban on sugary drinks
BERT-ED 10K	obama's court's ban in court	in new yorks's york taxs's taxs s
BERT-ED 10K + BERT aug	philippines's ban in philippines' ban in philippines' ban in philippines	in new york city, new york city's new york city's bans law
BERT-ED 10K + shuffling	philippines : lawmakers seek lawmaker's ban on lawmakers obama lawmakers arroyo's lawmakers arroyo's lawmakers	u. s. and new york's new new york city mayor' campaign campaign moves new york's mayor's campaign campaign
BERT-ED 10k + infilling	new new new new york city party party leader s. o. p. s. a. leader s. o. p.	philippines : s. o. p. to be suspended s. a. lawmakers s. ban s. a.'s
BERT-ED 10K + 2 tasks	president's decision to rebuke military law ends in conflict philippines arroyo's rebuke philippines's supreme court in	new yorkers face a challenge to soda industry in new yorkers in new yorkers' campaign campaign in new york city's

Table 4: Examples of English headlines generated by various models.

mance on English. According to numerical results in Table 3, the best augmentation method is BERT-based augmentation. However, as could be seen in Table 4, the outputs are still ungrammatical, though the meaning is closer to the true headlines. Similar results were obtained with other augmentation strategies.

Pretraining has a more positive effect, though repetitions and hallucinations are still possible, as can be seen in the last row in Table 4. Pretraining results in much longer output sequences, where in most of the cases only the first 5-6 words make sense, and then the model starts making repetitions as if it did not know where to stop. All BERT-ED-based models overuse possessive suffixes in an ungrammatical way. Text infilling strategy also results in overusing of abbreviations, though this problem disappears in a "2-task" pretraining (the last two rows in Table 4).

6. Conclusion

We investigated two systems for headline generation in a multilingual low-resource scenario. The first option is the employment of a pretrained multilingual encoder-decoder summarization model and the second one is combining two pretrained language models into an encoder-decoder architecture that is trained from scratch. We suggest that if the first option is available, i.e., there exists a pretrained multilingual NLG model for a specific low-resource language, it should be picked over the second one. The successful training of a randomly initialized cross-attention layer, connecting the two language models, is crucial for the model's performance and is dependent on a large corpus, such as the KPTimes train dataset containing round 260 K document. However, even in that scenario, the BERT-ED model is outperformed by an English BART model. While pretraining and data augmentation can drastically improve the performance of the BERT-ED models, it has little effect on the BART-based models, which have already been extensively pretrained on the same denoising tasks, text infilling and sentence shuffling, that we employ. Results also suggest that pretraining is a better option than data augmentation since the improvements are larger and since data augmentation had a negative effect on the performance of the BART-based models.

The best performance is achieved by the BART model trained on a large KPTimes train set. While this indicates that currently there is still no substitution for a large dataset, the BART model trained on the magnitudes smaller 10K dataset nevertheless still offers competitive performance. Scores that mBART achieves on the Estonian and Croatian datasets are lower, which could be caused by the fact that these languages are morphologically much richer languages than English. It might however also indicate that multilingual models cannot compete with the monolingual one, confirming the curse of multilinguality (Conneau et al., 2020).

On top of ROUGE-1, -2 and -L, we use two evaluation metrics that are less broadly used in the literature, measuring semantic similarity (SS) and sentence entailment (NLI). They are globally highly correlated with ROUGE scores, but allow for more fine-grained comparison when evaluating the impact of different augmentation and pretraining regimes.

The main focus of the future work will be on improving the quality of generated headlines in low-resource scenarios, by (1) introducing novel pretraining tasks and data augmentation techniques and by (2) pretraining monolingual encoder-decoder models on denoising tasks on as large corpora as can be obtained for low-resource languages.

7. Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 825153 (EMBEDDIA).

References

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dušek, O., Novikova, J., and Rieser, V. (2018). Findings of the E2E NLG challenge. *arXiv preprint arXiv:1810.01170*.

Dušek, O., Howcroft, D. M., and Rieser, V. (2019). Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426.

Gallina, Y., Boudin, F., and Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan, October–November. Association for Computational Linguistics.

Gkatzia, D. (2016). Content selection in data-to-text systems: A survey. *arXiv preprint 1610.08375*.

Graff, D., Kong, J., Chen, K., and Maeda, K. (2003). English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Higurashi, T., Kobayashi, H., Masuyama, T., and Murao, K. (2018). Extractive headline generation based on learning to rank for community question answering. In *COLING*, pages 1742–1753.

Huang, H., Liang, Y., Duan, N., Gong, M., Shou, L., Jiang, D., and Zhou, M. (2019). Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China, November. Association for Computational Linguistics.

Iwama, K. and Kano, Y. (2019). Multiple news headlines generation using page metadata. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 101–105, Tokyo, Japan, October–November. Association for Computational Linguistics.

Kaalep, H.-J., Muischnek, K., Uiboaed, K., and Veskis, K. (2010). The estonian reference corpus: Its composition and morphology-aware user interface. In *Baltic HLT*, pages 143–146.

Koloski, B., Pollak, S., Škrlj, B., and Martinc, M. (2021). Keyword extraction datasets for Croatian, Estonian, Latvian and Russian 1.0.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., et al. (2020). Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Liu, T., Li, H., Zhu, J., Zhang, J., and Zong, C. (2018). Review headline generation with user embedding. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 324–334. Springer.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ljubešić, N. and Erjavec, T. (2011). hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *EMNLP*, pages 1797–1807.

Purver, M., Pollak, S., Freienthal, L., Kuulmets, H.-A., Krustok, I., and Shekhar, R. (2021a). Ekspress news article archive (in estonian and russian) 1.0.

Purver, M., Shekhar, R., Pranjić, M., Pollak, S., and Martinc, M. (2021b). 24sata news article archive 1.0.

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. (2020). Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410.

Qi, W., Gong, Y., Yan, Y., Xu, C., Yao, B., Zhou, B., Cheng, B., Jiang, D., Chen, J., Zhang, R., et al. (2021a). Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006*.

Qi, W., Gong, Y., Yan, Y., Xu, C., Yao, B., Zhou, B., Cheng, B., Jiang, D., Chen, J., Zhang, R., Li, H., and Duan, N. (2021b). ProphetNet-X: Large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 232–239, Online, August. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Reiter, E. (2018). Hallucination in neural NLG. `https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/`. Accessed: 2020-03-02.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Shen, S., Zhao, Y., Liu, Z., Sun, M., et al. (2016). Neural headline generation with sentence-wise optimization. *arXiv preprint arXiv:1604.01904*.

Shen, S.-q., Chen, Y., Yang, C., Liu, Z.-y., Sun, M.-s., et al. (2018). Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ulčar, M. and Robnik-Šikonja, M. (2020). Finest bert and crosloengual bert: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Wen, T.-H., Gašić, M., Mrkšić, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September. Association for Computational Linguistics.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2020). Ccnet: Extracting high quality monolingual datasets from web crawl data. In *12th Conference on Language Resources and Evaluation (LREC 2020)*, page 4003–4012.

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China, November. Association for Computational Linguistics.